Opening the Black Box of Volatility: A Comparative Analysis of GARCH, RNN and Hybrid Architectures for S&P 500 Forecasting

Eddie Liu, Samuel Gu, Richard Li, Beihan Niu, Francis Xiao

February 2026

Abstract

Financial volatility forecasting is traditionally dominated by linear econometric models, which often struggle to capture the complex, non-linear dynamics of rapid regime shifts and volatility clustering. While Deep Learning architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks offer theoretical advantages in mapping these non-linearities, they are frequently criticized for their "black box" opacity and susceptibility to overfitting in low signal-to-noise financial environments. This paper presents a rigorous, leakage-safe empirical benchmark comparing three model families: a classical GARCH framework, pure end-to-end RNN architectures, and hybrid GARCH–RNN specifications. Using daily S&P 500 data from 2003 to 2024, we evaluate out-of-sample performance using both Mean Squared Error (MSE) and the asymmetric Quasi-Likelihood (QLIKE) loss. Crucially, we move beyond standard error metrics to conduct a forensic "Inside-Out" analysis of the recurrent networks. By extracting and tracking the internal Forget and Input gate activations during historical market shocks, we uncover how these neural networks dynamically manage memory retention during crises, thereby offering a novel framework for interpretable deep learning in algorithmic risk management.

1. Introduction

Momentum investing and trend-following strategies are well documented to generate significant returns during sustained bull markets, yet they remain structurally vulnerable to sudden, severe drawdowns during market panics—a phenomenon widely known as "momentum crashes". In previous research, we demonstrated that dynamically scaling portfolio exposure based on forecasted volatility could effectively "crash-proof" these strategies, substantially improving risk-adjusted returns and nearly doubling the Sharpe ratio compared to static weightings. Central to that framework was a hybrid modeling approach combining generalized autoregressive conditional heteroskedasticity (GARCH) with neural networks.

Despite the empirical success of hybrid volatility models in portfolio applications, the internal mechanics of the neural network component remain largely opaque. This lack of interpretability continues to hinder the broader adoption of deep learning methods in financial risk management. In particular, when a neural model successfully anticipates a volatility surge, it is often unclear which temporal patterns or regime signals the network is exploiting.

This study shifts the focus from portfolio application to methodological investigation. We conduct a leakage safe empirical comparison of recurrent neural network architectures, including LSTM and GRU variants, with classical GARCH models and hybrid GARCH–RNN specifications in the context of S&P 500 volatility forecasting.Using strictly expanding, non-overlapping training windows, we evaluate whether the additional complexity of gated recurrent architectures provides measurable benefits in the low signal-to-noise environment of daily financial returns.

Beyond forecast accuracy, we introduce an “inside-out” interpretability framework that extracts and analyzes gate activations during major market stress episodes. This approach provides new empirical evidence on how recurrent networks dynamically regulate memory in response to financial shocks. By linking internal network dynamics to observable market regimes, the paper contributes to the growing literature on interpretable deep learning in quantitative risk management and offers practical guidance for deploying neural volatility models in production settings.

2. Literature Review

The modeling of financial volatility has evolved significantly since the introduction of the ARCH and GARCH frameworks, which provided the first robust tools for capturing volatility clustering. Extensive empirical testing has repeatedly confirmed that the GARCH(1,1) specification, particularly when utilizing Student-t innovations to account for fat-tailed return distributions, serves as a highly effective and parsimonious baseline for broad equity indices like the S&P 50. In recent years, the literature has increasingly turned toward machine learning to capture the non-linear, asymmetric properties of financial markets. Recurrent Neural Networks, specifically LSTMs and GRUs, have shown promise due to their ability to maintain hidden states over time, theoretically allowing them to remember long-term volatility regimes. However, the direct application of end-to-end deep learning to raw financial returns frequently yields mixed results, largely due to the noisy nature of the data and the propensity of complex models to overfit. This challenge has given rise to hybrid methodologies. By combining linear models with non-linear neural networks, researchers attempt to leverage the strengths of both. Multiplicative or additive adjustments—where a neural network corrects the systematic bias of a baseline GARCH model—have been shown to yield more robust out-of-sample forecasts. Despite these advancements, the literature severely lacks studies that interrogate the internal gating mechanisms of these networks to verify if they are genuinely learning market regimes or merely memorizing recent noise.

3. Data and Experimental Setup

Data and Experimental Setup

The empirical analysis is conducted on the S&P 500 index using daily adjusted closing prices obtained from Yahoo Finance (via the yfinance API). The sample spans January 2003 through December 2024, covering multiple distinct volatility regimes, including the 2008 Global Financial Crisis, the 2018 Volmageddon episode, and the 2020 COVID-19 market shock.

Daily log returns are computed from adjusted prices. The primary forecasting target is the next-day squared log return, which serves as an unbiased, albeit noisy, proxy for the latent conditional variance. For descriptive and diagnostic purposes, we additionally compute a 21-trading-day rolling standard deviation of returns, annualized according to standard market conventions. This rolling measure is used strictly for visualization and regime context and is not used as the supervised learning target.

To ensure strict out-of-sample integrity and eliminate look-ahead bias, we employ an expanding rolling-window framework. The initial training window consists of 756 trading days (approximately three years), followed by a 252-day validation block. The out-of-sample test window is restricted to 21 trading days and advances forward in discrete, non-overlapping steps. All model refitting and hyperparameter selection occur strictly within each training–validation fold.

Transformation and Standardization Protocol

Financial return series and variance proxies exhibit strong heteroskedasticity and heavy-tailed behavior, which can destabilize neural network training and bias scale-sensitive loss functions. To address these issues while preserving leakage safety, all feature and target transformations are fitted exclusively on the training subset within each rolling split, and the resulting statistics are applied unchanged to the corresponding validation and test sets.

For positive variance-type targets (e.g., squared returns), we employ a log-standardization pipeline. Specifically, a small numerical offset is first added for stability, followed by a logarithmic transformation and standardization using training-set moments. This procedure reduces the impact of extreme volatility spikes and improves numerical conditioning for gradient-based optimization. For residual-based targets used in the Hybrid–Residual specification, standard (mean–variance) normalization is applied directly to the signed residual series. This distinction is important because residual targets are symmetric around zero, whereas variance targets are strictly positive and highly skewed.

Hybrid Model Inputs

Two hybrid architectures are considered.

- **Feature Hybrid models** incorporate the GARCH conditional variance forecast directly as an input feature to the neural network. In this specification, the network learns nonlinear corrections conditional on the GARCH volatility level.
- **Residual Hybrid models**, by contrast, are trained on the standardized residual innovations obtained after filtering returns through the GARCH model. This approach allows the recurrent network to model remaining nonlinear structure in the innovation process rather than the variance level itself.

Full implementation details of the transformation pipelines, including the log-standardization procedure and inverse mappings, are provided in the Methodology section.

4. Methodology

To isolate the marginal contribution of deep learning complexity, we construct a structured model zoo comprising six volatility forecasting architectures. All models generate one-step-ahead forecasts using a leakage safe expanding window protocol with monthly (21 trading day) refitting.

Let adjusted prices be Pt and define daily log returns:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right)$$

The primary evaluation target is next-day realized variance proxy:

$$RV_t = r_t^2$$

Baseline econometric model

The benchmark is a Student-t GARCH(1,1):

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \sim t_\nu(0,1)$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

The model generates the conditional variance forecast $\widehat{\sigma_{t|t-1^2}}$ with parameters re-estimated every 21 trading days.

Pure recurrent models

The pure neural models learn conditional variance directly from past returns. For a lookback window of length L, define the input sequence

$$\mathbf{x_t} = (r_{t-L}, \ldots, r_{t-1})$$

The network outputs log variance

$$h_t = f_\theta(\mathbf{x_t}), \qquad \sigma_{t|t-\widehat{1^2 = \mathrm{exp}}(h_t)}$$

which guarantees positivity.

Two backbone architectures are considered:

- **Pure LSTM**, which uses separate forget and input gates to manage long-term memory;
- **Pure GRU**, which replaces these with a single update gate and therefore contains fewer parameters, potentially improving generalization in noisy financial environments.

Feature hybrid models

Hybrid models incorporate the GARCH forecast as an econometric prior. Let

$$g_t = \sigma_{t|t-\widehat{1,\text{GARCH}}^2}$$

In the Feature Hybrid specification, the GARCH variance is appended to the return sequence:

$$\mathbf{x_t^{(F)}} = (r_{t-L}, \ldots, r_{t-1}, g_t)$$

The neural network then learns

$$\sigma_{t|t-\widehat{1^{2}=f_{\theta}}\left(\mathbf{x}_{\mathbf{t}}^{(\mathbf{F})}\right)}$$

We implement both Feature Hybrid LSTM and Feature Hybrid GRU, yielding two additional models.

Residual hybrid model

The Residual Hybrid instead decomposes realized variance into a GARCH baseline plus a nonlinear correction. Define the residual variance

$$u_t = RV_t - g_t$$

The recurrent network is trained to predict this residual:

$$\widehat{u_t} = f_\theta(\mathbf{x_t})$$

The final variance forecast is reconstructed as

$$\sigma_{t|t-\widehat{1^{2}=g}_{t}+\widehat{u_t}}$$

We implement this specification with both LSTM and GRU backbones.

Transformation and standardization

Financial variance proxies are highly skewed and heteroskedastic, which can destabilize neural training. To address this while maintaining strict leakage control, all transformations are fitted exclusively on the training subset within each rolling split.

For variance targets used in the pure and feature hybrid models, we apply log standardization:

$$y_t = \log(RV_t + \epsilon)\,, \qquad y_t^{std} = \frac{y_t - \mu_{\text{train}}}{\sigma_{\text{train}}}$$

After prediction, forecasts are mapped back via the inverse transform.

For the Residual Hybrid, the target u_t is signed and approximately symmetric, so standard mean-variance normalization is applied without the logarithmic step.

All reported out-of-sample forecasts are produced after inverse transformation and evaluated in variance space.

5. Evaluation Metrics

A common limitation in many machine learning applications to financial volatility forecasting is the default reliance on Mean Squared Error (MSE) as both the optimization and evaluation metric. Because financial returns exhibit heavy tails and pronounced volatility clustering, MSE can be overly sensitive to large but transient volatility realizations. Moreover, MSE is symmetric, penalizing over-prediction and under-prediction equally, which may be misaligned with the asymmetric risk considerations inherent in volatility forecasting.

Formally, the MSE of variance forecasts is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}\left(\sigma_t^2 - \widehat{\sigma_t^2}\right)^2$$

where σ_t^2 denotes the realized variance proxy and $\widehat{\sigma_t^2}$ is the one-step-ahead variance forecast. While widely used, this metric can be dominated by noise in squared returns and therefore may provide a distorted ranking of volatility models.

To address these issues, our evaluation framework places primary emphasis on the Quasi-Likelihood (QLIKE) loss function. QLIKE is specifically designed for strictly positive variance forecasts and is robust to noise in realized variance proxies. The metric is defined as:

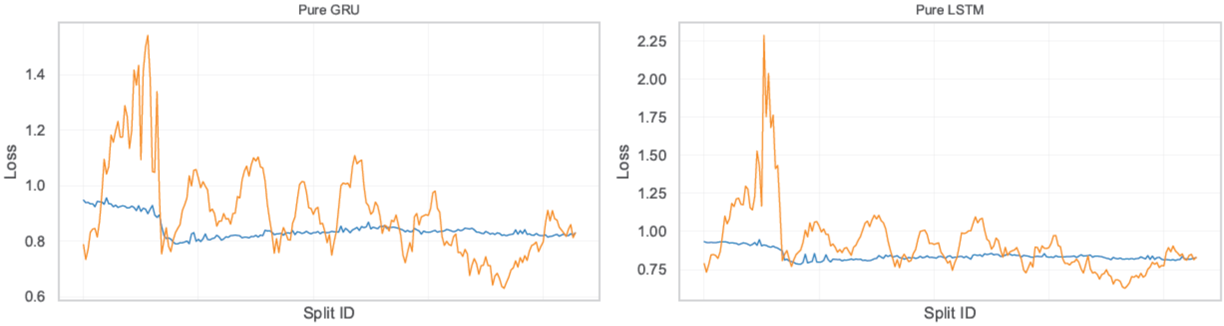
$$\text{QLIKE} = \frac{1}{N}\sum_{t=1}^{N}\left(\ln\left(\widehat{\sigma_t^2}\right) + \frac{\sigma_t^2}{\widehat{\sigma_t^2}}\right)$$

where σ_t^2 denotes the realized variance proxy and $\widehat{\sigma_t^2}$ is the one-step-ahead variance forecast.

Unlike MSE, which evaluates absolute squared deviations, QLIKE penalizes errors in relative variance scaling and arises directly from the Gaussian quasi-likelihood for conditional variance models. This property makes it particularly suitable for comparing volatility forecasts when the realized variance proxy is noisy, as is typical with daily squared returns. Accordingly, QLIKE serves as our primary model ranking criterion, while MSE is reported as a secondary diagnostic for completeness.

Overfit diagnoses

To assess the generalization behavior of the recurrent architectures, we examine the evolution of the best training and validation losses across rolling splits. Figure below plots the minimum loss achieved within each split for the Pure GRU and Pure LSTM models:



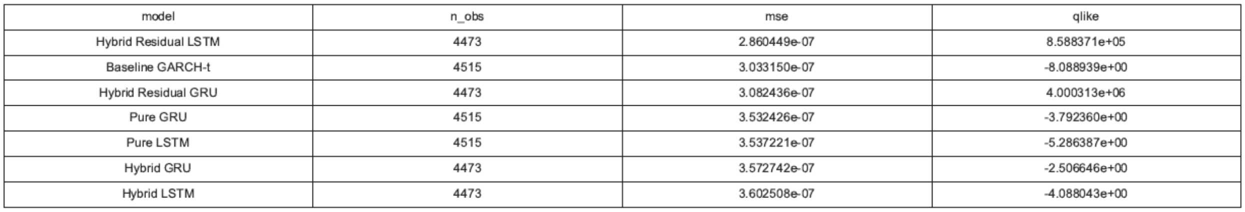
Several patterns emerge. First, both architectures exhibit elevated validation loss during early splits, coinciding with major market stress periods in the sample. This is expected given the heightened volatility of squared returns during crisis regimes and does not, by itself, indicate model overfitting. More importantly, the gap between training and validation loss remains relatively contained for both models throughout most of the sample. While the GRU occasionally achieves lower in-sample loss, its validation curve displays higher variability and more pronounced spikes, particularly in earlier splits. In contrast, the LSTM demonstrates more stable validation behavior and a smoother convergence profile over time. This difference is economically meaningful. The GRU's more aggressive parameter compression can lead to faster in-sample fitting but appears somewhat more sensitive to regime shifts in this low signal-to-noise environment. The LSTM's separate gating structure, while more parameter intensive, provides a modest but consistent improvement in out-of-sample stability.

Overall, the diagnostics do not indicate severe classical overfitting for either architecture. Instead, the results suggest that model risk in this setting arises primarily from variance miscalibration during high-volatility episodes rather than from excessive memorization of the training sample. This finding is consistent with the QLIKE-based ranking reported earlier, where models that appeared competitive under MSE exhibited substantial degradation once evaluated on variance-sensitive metrics.

6. Results

The out-of-sample expanding window backtest over 2003–2024 reveals a clear divergence between traditional squared-error evaluation and variance-appropriate scoring. Table I reports the full set of forecasting results across the model zoo.

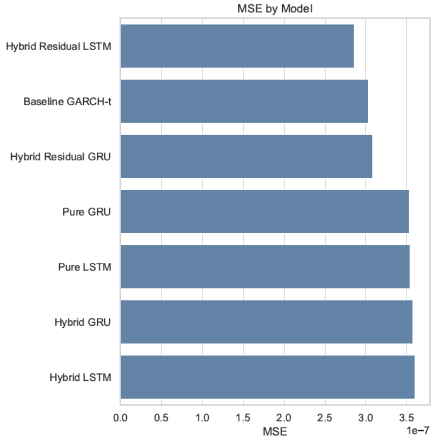
Table I:



model	n_obs	mse	qlike
Hybrid Residual LSTM	4473	2.860449e-07	8.588371e+05
Baseline GARCH-t	4515	3.033150e-07	-8.088939e+00
Hybrid Residual GRU	4473	3.082436e-07	4.000313e+06
Pure GRU	4515	3.532426e-07	-3.792360e+00
Pure LSTM	4515	3.537221e-07	-5.286387e+00
Hybrid GRU	4473	3.572742e-07	-2.506646e+00
Hybrid LSTM	4473	3.602508e-07	-4.088043e+00

Performance under MSE

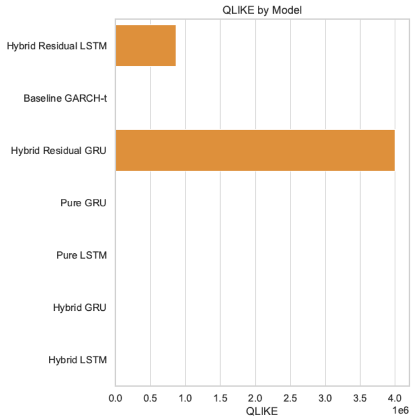
When evaluated using Mean Squared Error (MSE), the Residual Hybrid architectures appear to perform best. The Hybrid Residual LSTM achieves the lowest overall MSE (2.86×10^{-7}) marginally outperforming the baseline GARCH-t model (3.03×10^{-7}). Pure recurrent models and Feature Hybrid variants exhibit slightly higher but comparable MSE levels.

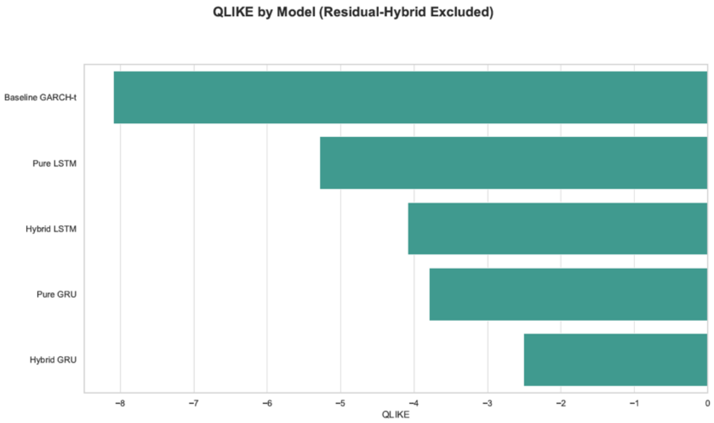


At face value, these results would suggest that modeling the residual variance component provides incremental predictive power. However, as discussed in Section 5, MSE is highly sensitive to the noise inherent in squared returns and may not reliably rank volatility models.

Performance under QLIKE

A markedly different picture emerges under the QLIKE metric. The Residual Hybrid models perform extremely poorly, generating large positive loss values (8.59×10^{5} for the LSTM variant and 4.00×10^{6} for the GRU). In contrast, the baseline GARCH-t establishes a strong performance floor with the best (most negative) QLIKE score of −8.09.





Among the neural architectures, the Pure LSTM delivers the most robust performance, achieving a QLIKE of −5.29 and outperforming both the Pure GRU (−3.79) and the Feature Hybrid LSTM (−4.09). The Feature Hybrid GRU shows further degradation (−2.51).

Interpretation

The stark reversal between MSE and QLIKE rankings highlights an important modeling distinction. The strong MSE performance of the Residual Hybrid models suggests that they fit the central mass of the return distribution effectively during low-volatility regimes. However, their extremely poor QLIKE scores indicate substantial miscalibration of variance scale, particularly during tail events. In several periods, these models produce near-zero or severely under-scaled variance forecasts, which QLIKE heavily penalizes due to its sensitivity to relative variance errors.

By contrast, the parsimonious GARCH-t model remains well calibrated across regimes, delivering the most reliable variance scaling. Among deep learning approaches, the Pure LSTM demonstrates the best overall robustness. Its gated memory structure appears sufficient to capture volatility dynamics directly from returns without requiring explicit econometric inputs.

Notably, augmenting the network with the GARCH conditional variance as an input feature does not improve out-of-sample performance and in some cases slightly degrades it. This suggests that, for daily S&P 500 data, the additional hybrid feature may introduce incremental noise rather than complementary signal.

7. Forensic Analysis: Unlocking the Gates

While the QLIKE results establish the statistical strength of the GARCH baseline, the internal dynamics of the recurrent networks provide additional insight into how deep learning models process volatility regimes. To better understand model behavior, we extract gate (lag1 – lag20) activations from the trained networks and examine their relationship with the VIX, a widely used proxy for market-implied volatility.

Gate semantics

In the LSTM architecture, the gates have distinct functional roles. The forget gate controls how much past information is retained, the input gate regulates the incorporation of new information into the cell state, and the output gate determines how much of the internal state is exposed to the next layer. The candidate gate proposes new state updates.

In contrast, the GRU combines memory retention and update decisions into a more compact structure via the reset and update gates, which reduces parameter count but also limits the model’s ability to decouple long-term memory management from short-term adaptation.

LSTM behavior across volatility regimes

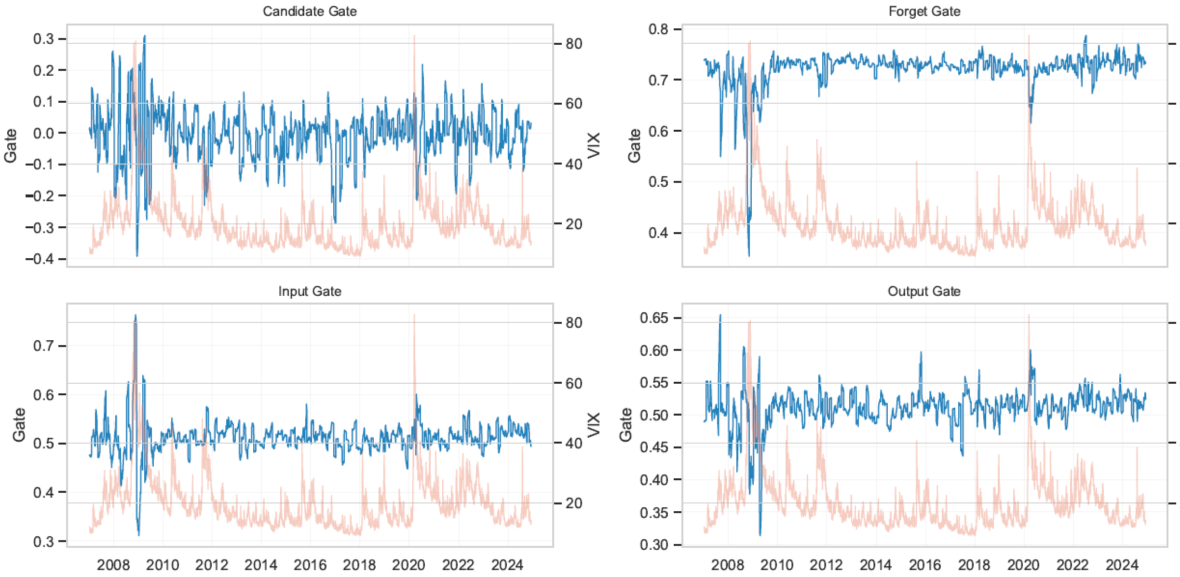
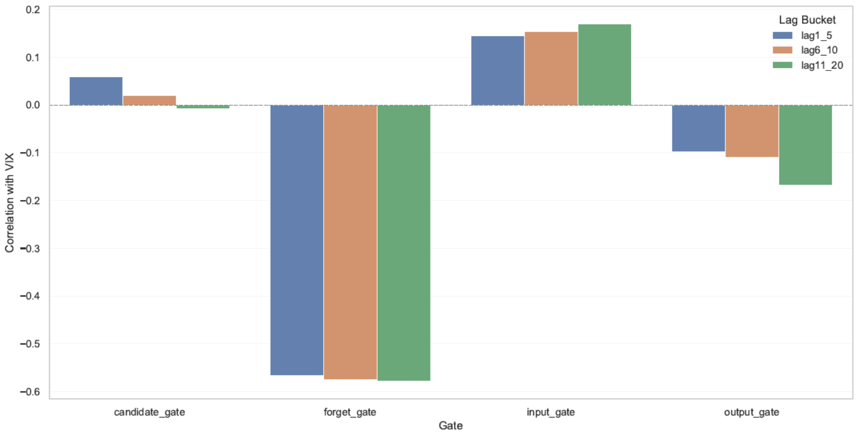


Figure above shows that the Pure LSTM exhibits clear regime sensitivity. The mean forget gate activation declines during high-volatility periods, indicating that the network reduces reliance on older historical information when market uncertainty rises. This pattern is consistent with adaptive memory management: during stress regimes, stale low-volatility history becomes less informative.

The bucket-level correlation analysis reinforces this interpretation. The forget gate shows a pronounced negative correlation with the VIX (approximately −0.57 at lag 1), which persists across longer lag aggregations. Conversely, the input gate exhibits a positive correlation with the VIX (approximately +0.14 to +0.17 depending on the lag bucket), suggesting that the network becomes more responsive to recent observations when volatility expectations increase as demonstrated in the figure below:



Taken together, these patterns indicate that the LSTM dynamically shifts from memory retention toward new information intake during periods of market stress. Importantly, this behavior aligns with the model's relatively strong out-of-sample QLIKE performance among the neural architectures.

Economic Interpretation

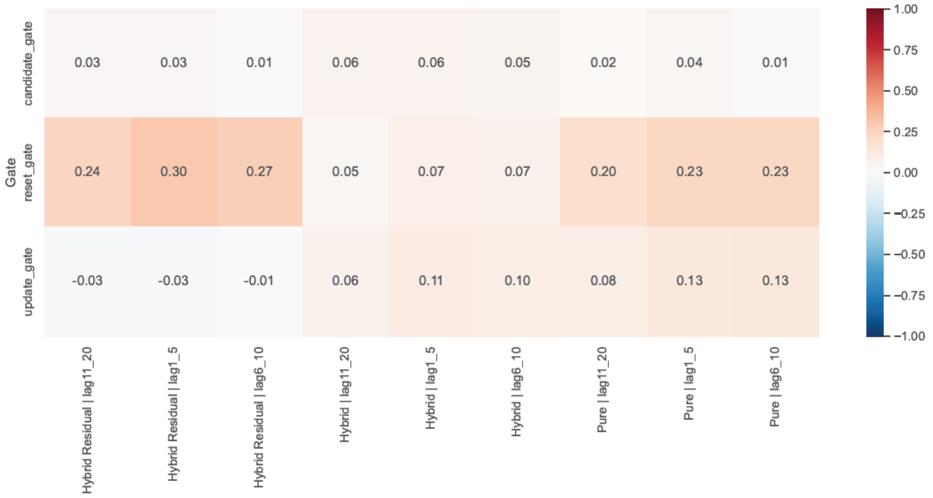
From an economic perspective, the observed negative relationship between the LSTM forget gate and the VIX is consistent with regime-adaptive learning behavior. During periods of elevated market stress, volatility dynamics are well known to undergo structural shifts, rendering long-horizon historical patterns less informative for short-term risk forecasting.

The LSTM appears to internalize this property. As the VIX rises, the average forget gate activation declines, implying that the network places less weight on distant historical states and effectively shortens its memory horizon. Economically, this behavior is intuitive: during volatility shocks, recent observations contain a higher proportion of relevant information about current risk conditions, while pre-crisis low-volatility regimes become progressively less predictive.

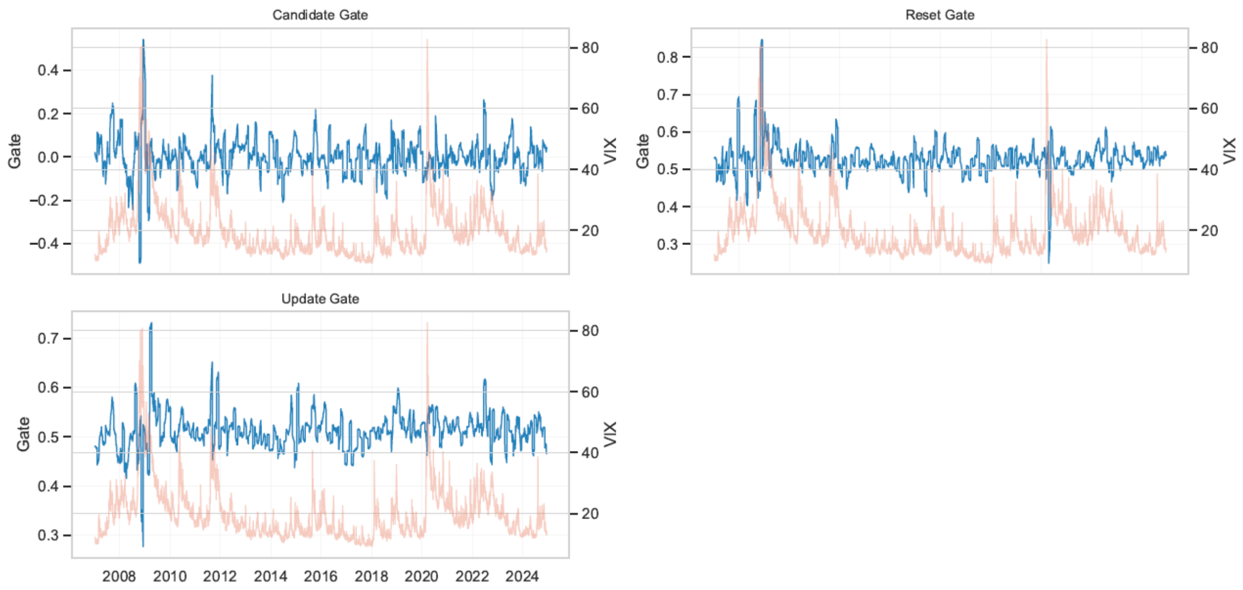
This adaptive memory contraction provides a plausible mechanism for the Pure LSTM's relatively strong QLIKE performance. By dynamically down-weighting stale historical context during high-volatility episodes, the model is better positioned to maintain appropriate variance scaling in rapidly changing market environments.

GRU behavior

The GRU models display markedly weaker regime adaptivity. As shown in Figure below, the reset and update gates exhibit 0.01 – 0.23 correlation with the VIX across most lag buckets.



Visually, the gate series remain comparatively stable even during major volatility spikes such as 2008 and 2020.



This muted response is consistent with the GRU's more compressed gating structure. Because the architecture does not fully decouple forgetting from state updating, it appears less able to dynamically reweight historical information in response to rapid volatility regime shifts. This likely contributes to the GRU's weaker QLIKE performance relative to the LSTM.

Linking gate dynamics to forecast performance

The forensic evidence suggests that the primary advantage of the LSTM architecture in this setting is not superior in-sample fit, but rather more effective regime-aware memory control. Models that successfully adjust their internal memory horizon during volatility shocks tend to maintain better variance calibration, which is heavily rewarded under QLIKE.

By contrast, architectures that insufficiently adapt their internal state—whether due to residual mis-specification or compressed gating—are more prone to variance under-scaling during tail events. This mechanism is consistent with the earlier finding that several neural variants performed adequately under MSE yet deteriorated sharply under QLIKE.

Summary of key findings

Across the full sample, three conclusions emerge:

1. The LSTM exhibits economically meaningful regime sensitivity, with the forget and input gates responding systematically to changes in market-implied volatility.
2. The GRU shows substantially weaker adaptive behavior, consistent with its more constrained gating structure.
3. Effective volatility forecasting in this low signal-to-noise environment appears to depend critically on dynamic memory management rather than purely on model flexibility or parameter count.

8. Limitation and Future research

Despite the encouraging insights, several avenues remain for further improvement.

Retraining Frequency

Due to computational constraints, the neural networks in this study are refit every 21 trading days, whereas the GARCH benchmark can be updated effectively at a daily frequency. A finer retraining schedule may allow the RNN models to respond more rapidly to regime shifts, potentially improving their QLIKE performance. Future work should investigate whether higher-frequency updating materially narrows the performance gap.

Direct optimization of QLIKE

The current neural architectures are trained primarily under squared-error objectives, while evaluation emphasizes QLIKE. Although this mirrors common practice, it introduces a potential objective mismatch. Training the networks directly under a QLIKE-consistent loss function may improve variance calibration, particularly during tail events. An interesting direction for future research is to examine how LSTM and GRU gating dynamics change when the models are explicitly optimized for QLIKE.

Richer realized volatility proxies

This study relies primarily on squared daily returns as a variance proxy. Incorporating high-frequency realized volatility measures may provide a cleaner supervision signal and potentially improve deep learning performance under QLIKE evaluation.

9. Conclusion

Overall, the empirical evidence suggests that while classical econometric models remain highly competitive for daily volatility forecasting, recurrent neural networks equipped with regime-adaptive memory mechanisms represent a promising direction for next-generation risk modeling.

The forensic gate analysis indicates that the primary advantage of the LSTM architecture arises from its ability to dynamically adjust its effective memory horizon in response to changing volatility regimes. Models that successfully modulate the balance between historical retention and new information intake appear better positioned to maintain variance calibration in the low signal-to-noise environment characteristic of equity markets.

Taken together, these findings suggest that future progress in neural volatility modeling may depend less on raw architectural complexity and more on the explicit design of state-dependent memory dynamics.

10. Appendix

Please see full report attached at the end.